EAFIT University (2022)

ARTICLE

Ensembled Nadaraya-Watson Regression using Robust Gaussian Kernel with Cross Validated Bandwidth

Alejandro Barrientos,[†] Alejandro Henao,[†] Juan Fernando Gallego,[†] Ricardo Morris,[†] and Juan Pablo Restrepo[*‡]

†MSc. in Data Science and Analytics, EAFIT University, Medellín, Colombia
‡MSc. in Applied Mathematics, EAFIT University, Medellín, Colombia
*Corresponding author. Email: jurest82@eafit.edu.co

Abstract

Non-parametric regression is an estimation of the relationship between variables without assuming any shape for the data. Methods based on Kernel Density Estimation such as the Nadaraya-Watson have shown errors when estimating out-of-sample data. In this paper, a method based on K-Fold Cross-Validation and Bagging is proposed to find an optimum bandwidth for the kernel estimation in order to reduce the test set error. This model was able to better capture the relationship between variables, with a maximum reduction of 99% in the test error, under simulated data with random noise.

Keywords: Non-Parametric Regression; Density Estimation; Bagging; Bandwidth Robustness

1. Introduction

Given two univariate random samples X and Y, the regression problem consists on estimating the expected value of a random variable Y, given a random variable X. This can be written as:

$$E(Y|X) = m(X) \tag{1}$$

where $m(X)$ is an unknown function. The most used method is known as Linear Regression (LR), which assumes a linear relationship between the variables, such that:

$$\widehat{Y} = \widehat{m}(X) = aX + b \tag{2}$$

where a represents the slope of the line and b the intercept with the axis represented by the random variable X. This model, besides the assumptions it requires, does not capture non-linear relationships effectively.

In order to overcome the assumptions of traditional LR, a non-parametric regression method can be implemented. In this case, we are focused on those derived from Kernel Density Estimation (KDE) methods, specifically the Nadaraya-Watson estimator. The model consists of a local weighted average using a kernel as the weighting function. It is then defined as (Nadaraya 1964; Watson 1964):

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h(\frac{x-x_i}{h})y_i}{\sum_{j=i}^{n} K_h(\frac{x-x_j}{h})} \tag{3}$$

where K_h is a kernel with a bandwidth h. This model has the advantage of being non-parametric, meaning it does not assume a distribution of the random variables, and not requiring to meet any assumptions. This allows the model to be able to capture most of the relationships between variables, even if they are non-linear and there's not a established function that correlates them.

One of the most used kernels for density estimation is the Gaussian Kernel, defined as:

$$K_h(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} \tag{4}$$

which is applied to each interval of size h. When using Gaussian-like kernels, it has been shown that the Scott rule for estimating the bandwidth produces outstanding results (Wasserman 2006). This bandwidth is calculated as $h = 1.06\sigma(x)n^{-\frac{1}{5}}$, and depends of the sample size and the variability of the data. It is demonstrated later in this paper, although this bandwidth is efficient, it usually introduces bias to the model, and misses to accurately explain the variability of the data, thus introducing the need for better bandwidth estimates.

Although generally the selection of the kernel is not crucial for the estimation, it is sometimes important depending on the shape of the data, especially with the presence of outliers. Some robust kernels have been developed, that are based on Gaussian-like shapes, with an extra smoothing of observations close to the tails of the kernel. A robust kernel estimator proposed by Wang et al. 2020 consists of a mixture of a regular-tailed kernel (Gaussian), and a kernel with a thick tail. This kernel in particular is defined as:

$$K_\omega(t) = \omega K_0(t) + (1 - \omega)\phi(t), \omega \in [0, 1] \tag{5}$$

where ω is the weight given to the kernel K_0 and $(1 - \omega)$ the weight given to ϕ. Here, ϕ represents the density function of the standard normal distribution, and K_0 is the thick-tail kernel defined as:

$$K_0(t) = \frac{1}{\sqrt{8\pi e}\Phi(1)} e^{-\frac{1}{2}\left(log(1+|t|)\right)^2} \tag{6}$$

Where $\Phi(1)$ is the cumulative distribution function of the standard normal distribution evaluated at one. As it will be shown, the Nadaraya-Watson estimator performs poorly in test sets and tends to overfit the data it is trained with. To reduce the variance of the test error, several methods based on resampling can be used. K-Fold Cross-Validation (KFCV) divides the sample in what's called folds, each one of approximately equal size. Each fold is used as the test sample in each iteration, for a total of k estimations of the test error. The average of the test error is then calculated, and reported as the Cross-Validation error. This methodology is often used to optimize hyperparameters since it manages directly the bias-variance trade-off (Hastie, Tibshirani, and Friedman 2009).

The testing error variance can be further reduced by selecting the training samples within the k-folds using bagging (Breiman 1994). After selecting one of the folds as the test set, resampling with replacement is done in the $k - 1$ folds left, using the methodology known as bootstrap. A number of r samples are generated, and each one of them used to fit the statistical model. It is then necessary to compute an estimation that represents the r samples as one, such as taking the element-wise mean or the median. This allows for a better estimation to be obtained, and therefore improving the explanation of the data behavior.

It is worth noting that these methodologies do not incur into issues related with overfitting since they better estimate the generalization error, that is, the error when estimating out-of-sample observations.

Since the Nadaraya-Watson estimator does not perform adequately in out-of-sample data, as it will be demonstrated throughout this work, it makes sense to apply the two methodologies above to improve the test error, since the variance and the bias are both reduced. In this paper, we propose a hyperparameter optimization for the proposed robust Gaussian-like kernel with bandwidth h, by using KFCV coupled with Bagging, in order to improve the out-of-sample regression performance. Obtained results are then compared with the Nadaraya-Watson non-parametric estimator that uses the Gaussian kernel and the Scott's rule bandwidth.

The paper is divided as follows: First the methodology section with an overview of the methods used and functions tested on, algorithm for h optimization and bagging steps. Following, results and discussion section showing comparison between the proposed model and the baseline. Finally, conclusions and references are drawn from the obtained results.

2. Methodology

An univariate sample is generated by simulation, using different generating functions, with a random white noise applied to them. The functions used were the following:

Function 1:

$$Y(X) = f(X) + 0.5f(3X + 0.23) + 0.5f(5X - 0.4) + 0.5f(7X - 2.09) + 0.5f(9X - 3) + 0.5N(0,1) \quad (7)$$

with $f(t) = \cos\left(\frac{\pi t}{15}\right)$.

Function 2:

$$Y(X) = 0.6X^5 + 10X^3 - 5X^2 + N(0, 1.5x10^7) \quad (8)$$

Function 3:

$$Y(X) = X^2 + f(X) + 0.5f(3X + 0.23) + 0.5f(5X - 0.4) + 0.5f(X - 2.09) + 0.5f(X - 3) + 0.5N(0,1) \quad (9)$$

with $f(t) = \sin\left(\frac{\pi t}{15}\right)$.

Function 4:

$$Y(X) = \log(X)\sin\left(\frac{\pi X}{15}\right) + N(0,1) \quad (10)$$

Function 5:

$$Y(X) = -0.01X^2 + 20\log(X+1) + 14\cos(X) + 10 + N(0,8) \quad (11)$$

Function 6:

$$Y(X) = \sqrt{X} + 10\sin(X) + N(0,1) \quad (12)$$

Two methods were compared in order to evaluate their accuracy in estimating out-of-sample data:

Algorithm 1 Estimate the optimal bandwidth h

```
Divide the dataset into training set and test set
Define a number $K$-folds to be used
Divide the training set into K folds
Define a closed interval $H$
Define the number of bootstrap samples $r$
Define kernel for the estimation as $K_h$
for $h$ in $H$ do
    for $k = 1$ to $K$ do
        Select the fold $k$ as the Cross-Validation test set
        Randomly choose $r$ samples with replacement from the remaining $k - 1$ folds
        Estimate $\widehat{m}_h(x)$ for each sample using $K_h$
        Calculate the element-wise mean of each estimation at $x$
        Calculate the error between the estimated values and the data in the $k$ fold
    end for
    Store $h$ and the mean of the errors between the $K$-folds
end for
Select the $h$ associated to the minimum error as the optimal bandwidth
```

1. Nadaraya-Watson estimator using Gaussian kernel and bandwidth calculated with the Scott's rule.
2. Nadaraya-Watson estimator using a robust kernel with optimal bandwidth estimated using KFCV and Bagging.

Algorithm 1 was used to estimate the optimum value of h in an interval H which minimized the Cross-Validation test error. This bandwidth was applied to the second method shown above, and later was compared to the one obtained with the first method. Results for each one of the functions were reported and analyzed over the simulated data.

The original dataset was split in two random subsets, the training set containing 70% of the total observations and the test subset containing the remaining 30%.

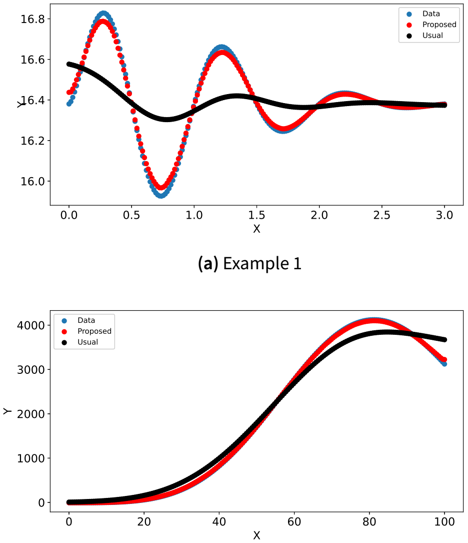
H interval was defined as a grid search around Scott's rule value, ranging between 0.05 and 1.5 times the Scott's base value for comparison, taking 40 steps inside the range. The number of models for the ensembled technique was 100, also defined for the r number of samples taken. K for the cross validated samples was 5, which is the default k-fold configuration and it proves useful while not being computationally expensive. The error used for comparison between cross validated folds and between methods was Mean Squared Error (MSE).

This paper will not focus on the effect of the weight (ω) given to the robust kernel in function (5), therefore its value will be fixed as 0.5 for all experiments. However, an additional run for *function*1 with $\omega = 1$ will be executed to generate new conclusions that might be useful as baseline for future works.

The methodology previously described was developed to improve the performance of the Nadaraya-Watson algorithm when making predictions y for x within the scope of the X values in the training set. Although, it is outside the horizon of this work, additional experiments with function 5 and 6 were made in order to evaluate and compare the performance of the models when trying to predict y for x values outside the scope of the training set.

3. Results and Discussion

An example on the regression results within a training set for the two methods described can be found in Figure 1a, being "Usual" the first method, and "Proposed" being the second method suggested in this paper. Due to the optimization of the bandwidth h, the regression model has a better fit to the data, with an error 70 times lower than the usual model. This model not only captures the variability



(a) Example 1

(b) Example 2

Figure 1. Non-Parametric regression using the two methods

of the data, but is also able to adjust its tails, which usually the Nadaraya-Watson estimator struggles to do.

Another comparison with different data can be seen in Figure 1b, which exhibits an error 123 times lower than the usual method. This shows that for even simpler dependency structure between variables, the proposed method is able to explain their behavior accurately. Here, the estimation of the tail to the right of the chart is also improved, thanks to the resampling performed that increases its density.

6 out-of-sample estimations can be seen in Figure 2, showing a better fit for the proposed model in all the evaluated functions. For the periodic data, the proposed model is able to explain the peaks, as seen in the central region of Figure 3a. Also, this regression method is able to follow the variability of the data, even when a higher amount of noise is present, which is the case of Figure 3e. This result is derived from the bagging that is made within the k-folds, generating samples with higher density of these observations for their estimation.

Errors associated to the out-of-sample estimation of each function are reported in Table 1. Overall, it can be seen that the hereby proposed model outperforms the usual estimation using the Gaussian kernel. This model better captures the relationships between the two variables, and reduces the test error. The minimum reduction obtained was of 51%, with a maximum of 99% obtained for the function 3. This clearly shows that the results from the usual Nadaraya-Watson estimator depends on the bandwidth used, and therefore it's necessary to evaluate different values to obtain better estimations.

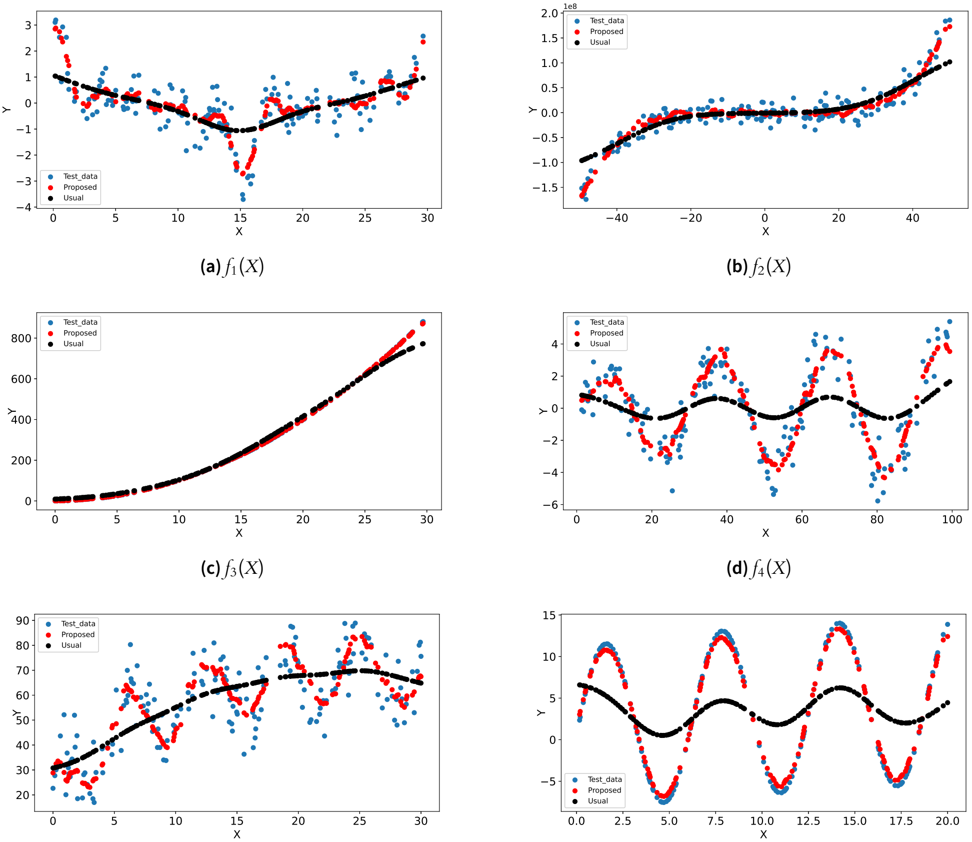
Besides the estimation of the curves, observations near the tails of the data are also improved, as can be seen in Figure 3b and 3c, which are often intervals with low density. This, on top of the bootstrapping, is possible due to the robust nature of the kernel selected.

Although the estimation is highly improved, the model still struggles to capture the variability of the data. However, given the fact that this estimation is made out-of-sample, the proposed method shows an outstanding performance when compared to the usual regression methods.

Figure 4 shows the estimations made for function 5 and 6 for x values that include the original

Table 1. Out-of-sample estimation error for the basic model and the proposed model

Experiment	Test mean square error		Error reduction
	Proposed model	Basic model	
1	0.28	0.63	56%
2	2.35 E+14	5.20 E+14	55%
3	2.60	503.24	99%
4	1.15	4.68	75%
5	68.63	140.96	51%
6	0.39	32.09	98%



(a) $f_1(X)$

(b) $f_2(X)$

(c) $f_3(X)$

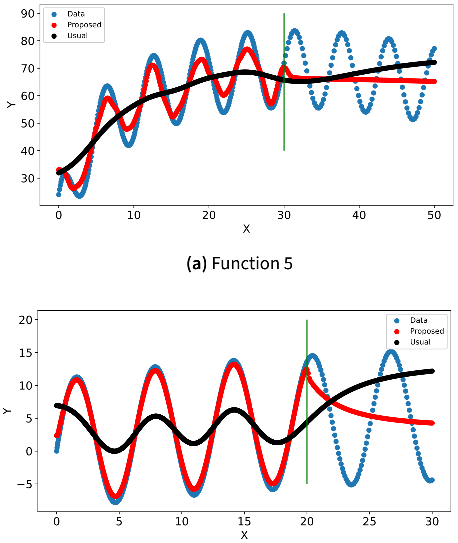
(d) $f_4(X)$

(e) $f_5(X)$

(f) $f_6(X)$

Figure 2. Out-of-sample estimates for the usual Gaussian kernel and the proposed model for each one of the 6 functions determined

scope of the training data and new observations with higher values of x that have not been observed by the model. The green vertical line separates the estimations for the observed x range from the extrapolated estimations. It is evident that both models perform poorly when trying to predict values outside the scope of the training set. However while the usual model always follows the final trend estimated for the known space, the proposed model seems to better fit the mean of the new data. This can be proved by looking at the MSE values.



(a) Function 5

(b) Function 6

Figure 4. Extrapolation for Function 5 and 6

Table 2 has the MSE for the estimation of the two functions in both the known x space and extrapolation range. Even though the estimation error is much higher when extrapolating than when making out-of-sample predictions within the training range, the proposed model always outperforms the usual in the scenarios studied.

Table 2. MSE for extrapolation in usual model and proposed model

Function	MSE for whole Data		MSE for extrapolation	
	Proposed	Usual	Proposed	Usual
$f_5(X)$	38.39	99.33	105.52	122.15
$f_6(X)$	7.09	35.79	47.95	85.13

Finally, as was mentioned before, the additional experiment for $function1$ with $\omega = 1$ in the equation 5 can be seen in figure 6. This means that the all weight is given to the robust kernel in the kernel estimator formula $K_\omega(t)$.

It's appreciated for this value, that there aren't significant differences for what was obtained with $\omega = 0.5$ in the same function 1. In this case, the MSE result for the proposed model in test data is 0.24, being slightly better than the previous one experiment: 0.28.

4. Conclusions

The results show that the proposed methodology outperforms the out-of-sample MSE for the usual Nadaraya-Watson model with Gaussian kernel using Scott's rule. Although the level of improvement varies from function to function it is consistently better in all cases.

The use of bagging with Nadaraya-Watson reduces the variability of the estimations and creates a steadier baseline to compare the MSE between all tested h values.

Since the optimal bandwidth selection is made by finding the lowest out-of-sample MSE through KFCV, the proposed methodology tends to overfit the outliers and has a reduced capacity to estimate

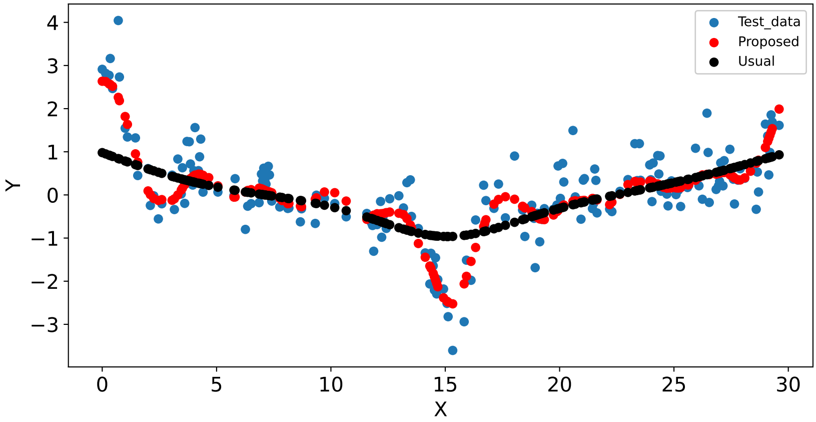


Figure 6. Function 1 ran with $\omega = 1$ in the robust kernel formula

the actual function behavior in presence of localized contamination. However, this flaw is also observed when applying the usual Nadaraya-Watson. It is proposed for future research to design an improved technique that minimizes the effect of contamination during the selection of the optimal h value. This could be carried out by identifying and removing outliers using Norm 2 and Convex-Hull based methods or any other statistical depth metric before Cross-Validation.

When changing the parameter ω in the kernel estimator equation 5, minimal differences were obtained for the MSE test data results, a Cross-Validation exercise for this parameter can further improve the performance of the proposed model slightly.

References

Breiman, Leo. 1994. Bagging predictors. *Machine Learning* 24 (2): 123–140. ISSN: 0885-6125. https://doi.org/10.1007/BF00058655.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning.* Springer New York. ISBN: 978-0-387-84857-0. https://doi.org/10.1007/978-0-387-84858-7.

Nadaraya, E. A. 1964. On estimating regression. *Theory of Probability & Its Applications* 9 (1): 141–142. https://doi.org/10.1137/1109020.

Wang, Shaoping, Ang Li, Kuangyu Wen, and Ximing Wu. 2020. Robust kernels for kernel density estimation. *Economics Letters* 191:109138. ISSN: 0165-1765. https://doi.org/https://doi.org/10.1016/j.econlet.2020.109138.

Wasserman, Larry. 2006. *All of nonparametric statistics (springer texts in statistics).* Berlin, Heidelberg: Springer-Verlag. ISBN: 0387251456.

Watson, Geoffrey S. 1964. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 26 (4): 359–372. ISSN: 0581572X, accessed May 1, 2022.